Decision Trees vs Neural Networks for Classification of Emotions from Text

Paolo Palumbo

September 10, 2024

1 Introduction

2 Data

The dataset used was a mix of the CrowdFlower dataset[?] and the Emotions dataset[?]. The CrowdFlower dataset contains 40,000 tweets labeled with 13 different emotions. The Emotions dataset contains 400,000 tweets labeled with 6 different emotions.

2.1 Exploratory data analysis and preparation

Some emotions that are present in the CrowdFlower dataset were remapped because they were too similar to other emotions and would make the classification task harder. The emotions and relative mappings are:

- Happiness
- Sadness
- Anger
- Worry
- Love
- Surprise
- Neutral
- Fun → Happiness
- Relief → Happiness
- Hate → Anger
- Empty → Neutral
- Enthusiasm → Happiness
- Boredom → Neutral

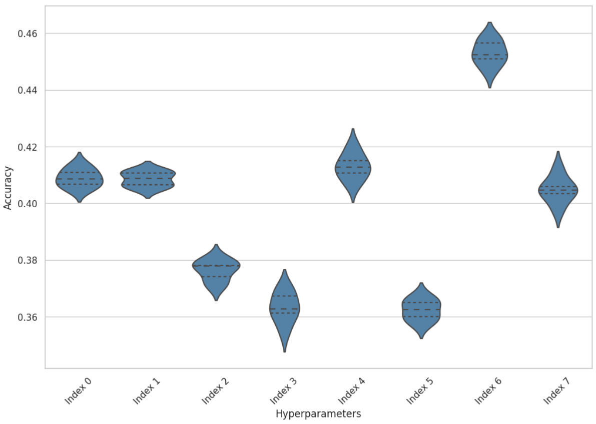

Figure 1: Class distribution of the full dataset

As shown in Figure 1, the dataset is imbalanced. The most common emotion is Happiness, and the least common is Neutral. This is because the larger dataset (Emotions) does not contain the Neutral class.

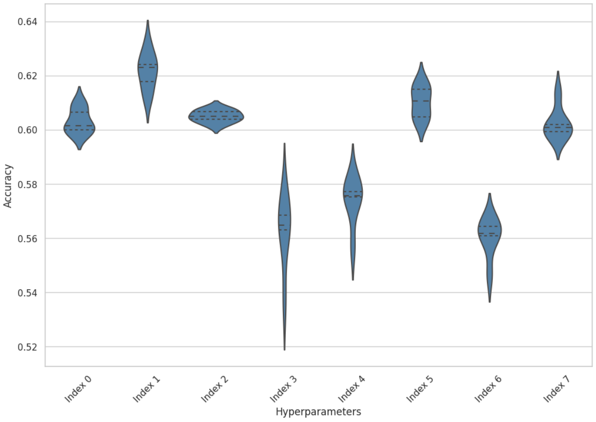

Figure 2: Text length distribution of the full dataset

The dataset was split into a training set and a test set with a 80% - 20% ratio.

2.2 Preprocessing

Multiple kinds of preprocessing were tested:

- *Bag of words (BoW)*
- *Word embeddings (WE)*
- *GloVe embeddings (GE)*

For the BoW and WE preprocessing, the text was first cleaned with a regular expression that removes urls, mentions and symbols. Then the text was tokenized and stemmed with the Snowball stemmer. Both **TFIDF** and **binary** encodings were tested for the BoW preprocessing. The GE preprocessing was done by applying the same cleaning regular expression but without stemming. The vectors used were the 6B 50d vectors from GloVe[**?**]. The main difference between the WE and GE preprocessing is that the WE vectors were trained on the dataset, while the GE vectors were pre-trained on a large corpus of text.

3 Models

Three different categories of models were considered:

- *Decision Trees*
- *Random Forests*
- *Neural Networks*
  - *Feedforward Networks*
  - *LSTM Networks*

Decision Trees and Random Forests both use the BoW preprocessing with **binary** encoding. Feedforward Neural Networks use the BoW preprocessing with **TFIDF** encoding. LSTM Networks were tested with both the WE and GE preprocessing.

3.1 Imbalanced Dataset

To handle the imbalanced dataset, a class weight was assigned to each class equal to $W_c = \frac{|D|}{|C||D_c|}$ where $|D|$ is the size of the dataset, $|C|$ is the number of classes and $|D_c|$ is the number of samples in class c. This weight was used in the criterion of the Decision Trees and Random Forests models, and in the loss function of the Neural Networks models.

3.2 Model Selection

To find a good hyperparameter configuration for each model, a random search was performed. The search was done with 10 iterations for each model (Decision Trees, Random Forests and Neural Networks). Each configuration was evaluated with a 6-fold cross-validation on the training set. Two subsequent random searches were performed, the second one narrowing the search space around the best configuration found in the first search. The search spaces for each model are shown in **??**, **??** and **??** for the first search, and in **??**, **??** and **??** for the second search.

Accuracy was used as the metric to evaluate the models.